Contrastive Context-Aware Learning for 3D High-Fidelity Mask Face Presentation Attack Detection

Ajian Liu, Chenxu Zhao, Zitong Yu, *Member, IEEE*, Jun Wan, *Senior Member, IEEE*, Anyang Su, Xing Liu, Zichang Tan, *Member, IEEE*, Sergio Escalera, *Senior Member, IEEE*, Junliang Xing, *Senior Member, IEEE*, Yanyan Liang, *Member, IEEE*, Guodong Guo, *Senior Member, IEEE*, Zhen Lei, *Senior Member, IEEE*, Stan Z. Li, *Fellow, IEEE*, and Du Zhang, *Senior Member, IEEE*

***Abstract*— Face presentation attack detection (PAD) is essential to secure face recognition systems primarily from high-fidelity mask attacks. Most existing 3D mask PAD benchmarks suffer from several drawbacks: 1) a limited number of mask identities, types of sensors, and a total number of videos; 2) low-fidelity quality of facial masks. Basic deep models and remote photoplethysmography (rPPG) methods achieved acceptable performance on these benchmarks but still far from the needs of practical scenarios. To bridge the gap to real-world applications, we introduce a large-scale High-Fidelity Mask dataset, namely HiFiMask. Specifically, a total amount of 54, 600 videos are recorded from 75 subjects with 225 realistic masks by 7 new kinds of sensors. Along with the dataset, we propose a novel Contrastive Context-aware Learning (CCL) framework. CCL is a new training methodology for supervised PAD tasks, which is able to learn by leveraging rich contexts accurately (e.g., subjects, mask material and lighting) among pairs of live faces and high-fidelity mask attacks. Extensive experimental evaluations on HiFiMask and three additional 3D mask datasets demonstrate the effectiveness of our method. The codes and dataset will be released soon.**

***Index Terms*— Face anti-spoofing, high-fidelity mask, contrastive context-aware learning.**

I. Introduction

FACE presentation attack detection (PAD) aims to secure a face recognition system from malicious presentation attacks (PAs), such as print attacks [1], video replay attacks [2], and 3D Mask attacks [3]. In recent years, face PAD approaches [4]–[10] for 2D attacks have made great progress, benefiting from the release of several large-scale, multi-modal and high-quality benchmark datasets [4], [11]–[13]. However, with the maturity of 3D printing technology, face mask has become a new type of PA to threaten face recognition systems' security. Compared with traditional 2D PAs, face masks are more realistic in terms of color, texture, and geometry structure, making it easy to fool a face PAD system designed based on coarse texture and facial depth information [4]. Fortunately, some works have been devoted to 3D mask attacks, including design of datasets [14]–[18] and algorithms [19]–[23].

In terms of the composition of 3D mask datasets, several drawbacks limit the generalization ability of data-driven algorithms. From existing 3D mask datasets shown in Tab. I, one can see some of these drawbacks: (1) **Bias of identity**. The number of mask subjects is less than the number of real face subjects. Even for some public datasets as [17], [20], [28], [29], the mask and live subjects correspond to completely different identities, which may produce the model to mistake identity as a discriminative PAD-related feature; (2) **Limited number of subjects and low skin tone variability**. Most datasets contain less than 50 subjects, with low or unspecified skin tone variability; (3) **Limited diversity of mask materials**. Most datasets [14], [22], [26]–[30] provide less than 3 mask materials, which makes it difficult to cover the attack masks that attackers may use; (4) **Few scene settings**. Most datasets [14], [26], [27], [30] only consider single deployment scenarios, without covering complex real-world scenarios; (5) **Controlled lighting environment**. Lighting changes pose a great challenge to the stability of rPPG-based PAD methods [25]. However, all existing mask datasets avoid this by setting the lighting to a fixed value, *i.e.*, daylight, office light; (6) **Obsolete acquisition devices**. Many datasets use outdated acquisition devices regarding the resolution and imaging quality. To alleviate previous issues, we introduce a large-scale 3D High-Fidelity Mask dataset for face PAD, namely **HiFiMask**. As shown in Tab. I, HiFiMask provides 25 subjects with yellow, white, and black skin tones to facilitate fair artificial intelligence (AI) and alleviate skin-caused biases (a total of 75 subjects). Each subject provides 3 kinds of high-fidelity masks with different materials (*i.e.*, plaster, resin, and transparent). Thus, a total of 225 masks are collected. In terms of recording scenarios, we consider

Manuscript received 7 December 2021; revised 12 April 2022; accepted 7 May 2022. Date of publication 4 July 2022; date of current version 13 July 2022. This work was supported in part by the National Key Research and Development Plan under Grant 2021YFF0602103, in part by the External Cooperation Key Project of Chinese Academy of Sciences (173211KYSB20200002), in part by the Chinese National Natural Science Foundation under Project 61876179 and Project 61961160704, in part by the Science and Technology Development Fund of Macau under Project 0070/2020/AMJ, in part by the Open Research Projects of Zhejiang Laboratory under Project 2021KH0AB07, in part by the Spanish Project (PID2019-105093GB-I00), in part by the Institución Catalana de Investigación y Estudios Avanzados (ICREA) under the ICREA Academia Program, and in part by the InnoHK Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Walter J. Scheirer. *(Ajian Liu, Chenxu Zhao, and Zitong Yu are co-first authors.) (Corresponding author: Jun Wan.)*

Please see the Acknowledgment section of this article for the author affiliations.

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2022.3188149, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2022.3188149

TABLE I

COMPARISON OF THE PUBLIC 3D FACE ANTI-SPOOFING DATASETS. 'Y,' 'W,' AND 'B' ARE SHORTHAND FOR YELLOW, WHITE, AND BLACK, RESPECTIVELY. 'SUB.', 'MASK ID.' AND 'LIGHT. COND.' DENOTE 'LIVE SUBJECTS,' 'MASK IDENTITY NUMBERS,' AND 'LIGHTING CONDITION,' RESPECTIVELY. NUMBER WITH '*' DENOTES THIS NUMBER IS STATISTICALLY INFERRED AND THERE MAY BE INACCURACIES

Dataset, Year	Skin tone	#Sub.	#Mask Id.	Material	Scenes	Light. Cond.	Devices	#Videos (#Live/#Fake)
3DMAD, [14]	W/B	17	17	Paper, hard resin	Controlled	Adjustment	Kinect	255(170/85)
3DFS-DB, [27]	Y	26	26	Plastic	Office	Adjustment	Kinect, Carmine 1.09	520(260/260)
BRSU, [20]	Y/W/B	137	6	Silicone, Plastic Resin, Latex	Disguise Counterfeiting	Adjustment.	SWIR, Color	141(0/141)
MARsV2, [26]	Y	12	12	ThatsMyFace REAL-F	Office	Room light, Low light Bright light, Warm light Side light, Up side light	Logitech C920, EOS M3, Nexus 5, iPhone 6 Samsung S7, Sony Tablet S	1008 (504/504)
SMAD, [28]	-	Online	Online	Silicone	-	varying lighting	Varying cam.	130(65/65)
MLFP, [29]	W/B	10	7	Latex, Paper	Indoor, Outdoor	Daylight	Visible Near infrared, Thermal	1350 (150/1200)
ERPA, [30]	W/B	5	6	Resin, Silicone	Indoor	Room light	Xenics Gobi, thermal cam. Intel Realsense SR300	86
WMCA, [17]	Y/W/B	72	7*	Plastic Silicone, Paper	Indoor	Office light LED lamps, day-light	Intel RealSense SR 300 Seek Thermal, Compact PRO.	1679 (347/1332)
CASIA-SURF 3DMask, [22]	Y	48	48	Plaster	Indoor, Outdoor	Normal light, Back light Front light, Side light Sunlight, Shadow	Apple, Huawei Samsung	1152 (288/864)
HiFiMask (Ours), 2021	Y/W/B	75	75	Transparent Plaster, Resin	White, Green Tricolor, Sunshine Shadow, Motion	NormalLight, DimLight BrightLight, BackLight SideLight, TopLight	iPhone11, iPhoneX MI10, P40, S20 Vivo, HJIM	**54,600** (13,650/40,950)

6 scenes, including indoor and outdoor environments with extra 6 directional and periodic lighting. As for the sensors for video recording, 7 mainstream imaging devices are used. In total, we collected 54, 600 videos, of which the live and mask videos are 13, 650 and 40, 950, respectively.

For 3D face PAD approaches, both appearance-based [17], [22], [31], [32] and remote photoplethysmography (rPPG)-based [21], [33], [34] methods have been developed. As illustrated in Fig. 1, although both appearance-based method ResNet50 [24] and rPPG-based method GrPPG [25] perform well on 3DMAD [14] and HKBU-MARs V2 (briefly named MARsV2) [26] datasets, these methods fail to achieve high performance on the proposed HiFiMask dataset. On the one hand, the high-fidelity appearance of 3D masks makes it harder to be distinguished from the bonafide. On the other hand, temporal light interference leads to pseudo 'liveness' cues for even 3D masks, which might confuse the rPPG-based attack detector. To tackle the challenges about high-fidelity appearance and temporal light interference, we propose a novel Contrastive Context-aware Learning framework, namely **CCL**, which learns discriminability by comparing image pairs with diverse contexts. Various kinds of image pairs are organized according to the context attribute types, which provide rich and meaningful contextual cues for representation learning. For instance, constructing face pairs from the same identify with both bonafide (*i.e.*, skin material) and mask presentation (*i.e.*, resin material) could benefit the fine-grained material features learning. Due to the significant appearance variations between some 'hard' positive pairs, the proposed CCL framework's convergence might sometimes be unstable. To alleviate the influence of such 'outlier' pairs and accelerate convergence, the Context Guided Dropout module, namely **CGD**, is proposed for robust contrastive learning via adaptively discarding parts of unnecessary embedding features. Our main contributions are summarized as follows:

- A large-scale 3D high-fidelity mask face PAD dataset named HiFiMask is released. Compared with public 3D mask datasets, HiFiMask has several advantages, such as realistic masks and amount of data in the term of identities, sensors and videos. Specifically, It consists of 54, 600 videos, 75 subjects with 3 kinds of high-fidelity masks, which is larger at least 16 times than the existing datasets in terms of data amount.
- We propose a novel Contrastive Context-aware Learning (CCL) framework to efficiently leverage rich and fine-grained context between live and mask faces for discriminative feature representation.
- Extensive experiments conducted on the HiFiMask and three other public 3D mask datasets demonstrate the challenges of HiFiMask and the effectiveness of the proposed method.

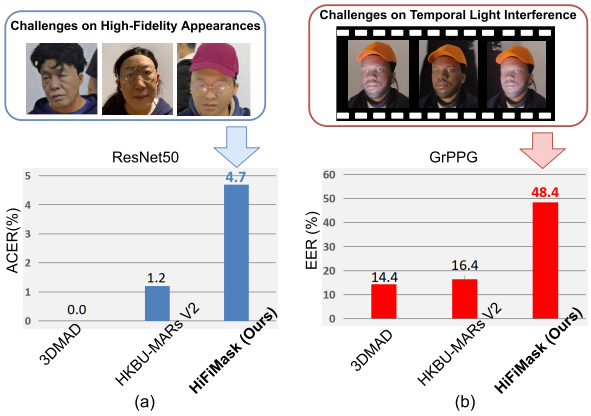

Fig. 1. Performance of ResNet50 [24] and GrPPG [25] on 3DMAD [14], MARsV2 [26] and our proposed HiFiMask datasets. Despite satisfying mask PAD performance on 3DMAD [14] and MARsV2 [26], these methods fail to achieve convincing results on HiFiMask. (a) Results on appearance changes. (b) Results on temporal light changes.

II. RELATED WORK

A. 3D Mask Datasets

Recently, several 3D mask face PAD datasets have been released. As listed in Tab. I, 3DMAD [14] is the first publicly available 3D mask dataset, which consists of 255 videos from 17 subjects, and the masks are made of paper and hard resin. Subsequent datasets 3DFS-DB [27], HKBU-MARs V2 [26], and BRSU Skin/Face/Spoof (briefly named BRSU) [20] improve previous drawbacks in terms of acquisition devices, mask types, and lighting environment. The

recent CASIA-SURF 3DMask (briefly named 3DMask) [22] has a large number of videos under various lighting conditions using various recording sensors. Still, it has a limited number of subjects and mask types. Besides common RGB modality, several multi-modal mask datasets such as MLFP [29], ERPA [30], and WMCA [17] extend the study from visible light to near-infrared and thermal spectrums. Overall, there are three main limitations of existing 3D mask datasets: 1) a limited number of samples, resulting in potential overfitting; 2) lack of clear attribute information (*e.g.*, skin tone and lighting) for evaluating the impact of external factors; and 3) the masks are not realistic enough in terms of color texture and structure, and they are recorded under stable lighting conditions.

B. Face PAD Approaches Based on 3D Mask Datasets

Compared with 2D presentation attacks, 3D mask attacks are more realistic to live faces in terms of depth shape and color texture. On the temporal side, several rPPG-based methods [21], [25], [34], [35] are proposed according to the evidence that periodic rPPG pulse cues could be recovered from the live faces but noisy for the mask attacks. Li *et al.* [25] was the first to leverage the facial rPPG signals' frequency statistics for mask attacks detection. Liu *et al.* [21], [34], [35] combined both local rPPG signals and global background noises to learn consistent rPPG features for mask PAD.

As for metric learning-based PAD approaches, contrastive loss [36] and triplet loss [37], [38] are utilized to widen the distance between the live faces and PAs. Recently, contrastive learning [39]–[42] achieved outstanding performance in self-supervised generic object classification. In [43], supervised contrastive learning is proposed for boosting performance upon using basic cross-entropy loss. Despite with similar design philosophy, the proposed CCL is different from [43] in both data pair generation and dropout regularization steps.

The approaches as mentioned above might be unreliable under the following situations: 1) high-fidelity mask attack with realistic appearance; 2) dynamic light flashing to disturb rPPG recovery; 3) metric learning-based constraints obtain unsatisfactory performance in PAD tasks; and 4) existing self-supervised or supervised contrastive learning approaches are not suitable for fine-grained binary classification task like 3D mask PAD. To tackle these issues, we propose a contrastive context-aware learning framework to explicitly mine the discriminative features among bonafide/mask appearance and complex scenarios.

III. HiFiMask Dataset

Given the shortcomings of the current mask datasets, we carefully designed and collected a HiFiMask dataset, which provides 5 main advantages over previous existing datasets. **Advantage 1:** To the best of our knowledge, HiFiMask is currently the largest 3D face mask PAD dataset, which contains 54, 600 videos captured from 75 subjects of three skin tones, including 25 subjects in yellow, white, and black, respectively. **Advantage 2:** HiFiMask provides 3 high-fidelity masks with the same identity, which are made of transparent, plaster, and resin materials, respectively. As shown in Fig. 2, our realistic masks are visually difficult to be distinguished from live faces. **Advantage 3:** We consider 6 complex scenes, *i.e.*, White Light, Green Light, Periodic Three-color Light, Outdoor Sunshine, Outdoor Shadow, and Motion Blur for video recording. Among them, there is periodic lighting within [0.7, 4]Hz for the first three scenarios to mimic the human heartbeat pulse, thus might interfere with the rPPG-based mask detection technology [25]. Please see Sec. V-A for corresponding rPPG analysis. **Advantage 4:** We repeatedly shoot 6 videos under different lighting directions (*i.e.*, NormalLight, DimLight, BrightLight, BackLight, SideLight, and TopLight) for each scene to explore the impact of directional lighting. **Advantage 5:** 7 mainstream imaging devices (*i.e.*, iPhone11, iPhoneX, MI10, P40, S20, Vivo, and HJIM) are utilized for video recording to ensure high resolution and imaging quality.

Fig. 2. Samples from the HiFiMask dataset. The first row shows 6 kinds of imaging sensors. The second row shows 6 kinds of appendages, among which E, H, S, W, G, and B are the abbreviations of Empty, Hat, Sunglasses, Wig, Glasses, and messy Background, respectively. The third row shows 6 kinds of illuminations, and the fourth row represents 6 deployment scenarios.

A. Acquisition Details of HiFiMask

Here we review the HiFiMask acquisition details in terms of equipment preparation, collection rules, and data preprocessing.

1) Equipment Preparation: In order to avoid identity information to interfere with the algorithm design, the plaster, transparent and resin masks are customized for real people. We use pulse oximeter CMS60C to record real-time Blood Volume Pulse (BVP) signals and instantaneous heart rate data from live videos. For scenes of White light, Green light, Periodic Three-color light (Red, Green, Blue and their various combinations), we use a colorful lighting to set the periodic frequency of illumination changes which is consistent with the rang of human heart rate. The change frequency is randomly set between [0.7,4]Hz and recorded for future research. At the same time, we use an additional light source to supplement the light from 6 directions (NormalLight, DimLight, BrightLight, BackLight, SideLight, and TopLight). The light intensity is randomly set between 400-1600 lux.

2) Collection Rules: To improve the video quality, we pay attention to the following steps during the acquisition process: 1) All masks are worn on the face of a real person and uniformly dressed to avoid the algorithm looking for clues outside the head area; 2) Collectors were asked to sit in front of the acquisition system and look towards the sensors with small head movements; 3) During data collection stage, a pedestrian

was arranged to walk around in the background to interfere with the algorithm to compensate the reflected light clues from the background [21]; 4) All live faces or masks were randomly equipped with decorations, such as sunglasses, wigs, ordinary glasses, hats of different colors, to simulate users in a real environment.

3) Data Pre-Processing: In order to save storage space, we remove irrelevant background areas from original videos, such as the part below the neck. As shown in Fig. 2, the reserved face area is obtained through the following steps. For each video, we first use Dlib [44] to detect the face in each frame and save its coordinates. Then find the largest box from all the frames of in videos to crop the face area. After face detection, we sample 10 frames at equal intervals from each video. Finally, we name the folder of this video according with the following rule: *Skin_Subject_Type_Scene_Light_Sensor*. Note that for the rPPG baseline [25], we use the first 10-second frames of each video for rPPG signal recovery without frame downsampling.

To expose the realness of masks in our proposed HiFiMask dataset, we calculate the similarity between a real face and its corresponding mask within three popular mask datasets. As shown in Fig. 3, the similarity calculation is conducted by FaceX-Zoo [45] and InsightFace [46]. By sampling some typical examples, we find that the similarity in HiFiMask is notably higher than in MARsV2 [26] and 3DMask [22].

In ***Appendix***, we show some samples of one subject with yellow skin tone. Six modules with different background lighting colors represent 6 kinds of scenes including white, green, three-color, sunshine, shadow and motion. The top of each module is the sample label or mask type, and the bottom is the scene type. Each row in one module corresponds to 7 types of imaging sensors (one frame is randomly selected for each video), and each column shows 6 kinds of lights.

B. Evaluation Protocol and Statistics

We define three protocols on HiFiMask for evaluation: Protocol 1-'seen', Protocol 2-'unseen' and Protocol 3-'openset'. The information used in the corresponding protocol is described in Tab. II.

1) Protocol 1-'Seen': Protocol 1 is designed to evaluate algorithms' performance when the mask types have been 'seen' in training and development sets. In this protocol, all skin tones, mask types, scenes, lightings, and imaging devices are presented in the training, development, and testing subsets, as shown in the second and third columns of Protocol 1 in Tab. II.

2) Protocol 2-'Unseen': Protocol 2 evaluates the generalization performance of the algorithms for 'unseen' mask types. Specifically, we further define three leave-one-type-out testing subprotocols based on Protocol 1 to evaluate the algorithm's generalization performance for transparent, plaster, and resin mask, respectively. For each protocol that is shown in the fourth columns of Protocol 2 in Tab. II, we train a model with 2 types of masks and test on the left 1 mask. Note that the 'unseen' protocol is more challenging as the testing set's mask type is unseen in the training and development sets.

3) Protocol 3-'Openset': Protocol 3 evaluates both discrimination and generalization ability of the algorithm under the open set scenarios. In other words, the training and developing sets contain only parts of common mask types and scenarios while there are more general mask types and scenarios on testing set. As shown in Tab. II, based on Protocol 1, we define training and development sets with parts of representative samples while full testing set is used. Thus, the distribution of testing set is more complicated than the training and development sets in terms of mask types, scenes, lighting, and imaging devices. Different from Protocol 2 with only 'unseen' mask types, Protocol 3 considers both 'seen' and 'unseen' domains as well as mask types, which is more general and valuable for real-world deployment.

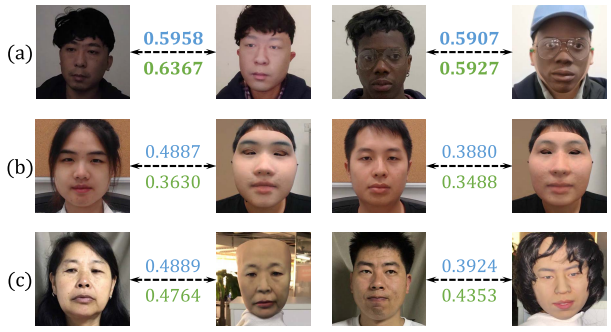

Fig. 3. The similarity of real faces and masks from different datasets. (a), Images from our HiFiMask. (b), Images from MARsV2. (c), Images from 3DMask. Results of FaceX-Zoo is marked in blue color and InsightFace in green. **Best viewed in color**.

TABLE II

STATISTICS OF EACH PROTOCOL IN HIFIMASK. PLEASE NOTE THAT PROTOCOLS 1, 2, AND 3 IN THE FOURTH COLUMN INDICATE TRANSPARENT, PLASTER, AND RESIN MASK, RESPECTIVELY

Pro.	Subset	subject	Masks	# live	# mask	# all
1	Train	45	1&2&3	8,108	24,406	32,514
	Dev	6	1&2&3	1,084	3,263	4,347
	Test	24	1&2&3	4,335	13,027	17,362
2_1	Train	45	2&3	8,108	16,315	24,423
	Dev	6	2&3	1,084	2,180	3,264
	Test	24	1	4,335	4,326	8,661
2_2	Train	45	1&3	8,108	16,264	24,372
	Dev	6	1&3	1,084	2,174	3,258
	Test	24	2	4,335	4,350	8,685
2_3	Train	45	1&2	8,108	16,233	24,341
	Dev	6	1&2	1,084	2,172	3,256
	Test	24	3	4,335	4,351	8,686
3	Train	45	1&3	1,610	2,105	3,715
	Dev	6	1&3	210	320	536
	Test	24	1&2&3	4,335	13,027	17,362

IV. METHODOLOGY

In this section we introduce the Contrastive Context-Aware Learning (CCL) framework for 3D high-fidelity mask PAD. CCL train models by contrastive learning meanwhile in a supervised learning manner. As illustrated in Fig. 4, CCL contains a data pair generation module to generate input data by leveraging rich contextual cues, a well-designed contrastive learning architecture for face PAD tasks, and the Context Guided Dropout (CGD) module accelerates the network convergence during the early training stages. The complete training procedure of CCL is described in Alg. 1.

A. Data Pair Generation

To effectively leverage rich contextual cues (*e.g.*, skin, subject, type, scene, light, sensor, and inter-frame information)

in HiFiMask datasets, we organize the data into pairs and freeze some of the contexts to mine the discrimination of other contexts, *e.g.*, we select a live face and a resin mask face from the same subject. Then the contrast is the discrepancy between the material of the live skin and resin. We split the live and mask faces into fine-grained patterns to generate a variety of meaningful contextual pairs. As shown in Tab. IV, we generate contextual pairs in the following way: 1) in Pattern. 1, we sample two frames from one single video as one kind of positive context pair; 2) in Pattern. 5, we sample one fine-grained mask category and the living category with the same subject as one negative context pair; 3) the positive and negative context pairs, including but not limited to the above combinations, are generated as the training set. The diagrammatic sketch of which contexts are compared in each pattern can be found in the left part of Fig. 4. The ablation of the pattern generation is studied in Sec. V-A.

B. Network Architecture

Recently, self-supervised contrastive learning, such as SimCLR [40], BYOL [41], and SimSiam [42], achieved outstanding performance in downstream prediction tasks. The purpose of these algorithms is to learn effective visual representations in advance. Therefore, taking the FAS task as a downstream task for the first time, we consider building the approach on a self-supervised contrastive learning framework, which aims to learn useful visual representations in advance. Inspired by the architectures in self-supervised learning [40]–[42], we extend the self-supervised batch contrastive approach to the fully-supervised setting, allowing us to effectively leverage label information, and propose the CCL framework, consisting of an online network and a target network for pairwise contrastive information propagation. At the same time, an extra classifier is used for explicit supervision. As shown in Fig. 4, well-organized contextual image pairs are utilized as the inputs of the CCL. The inputs are sent to the *online* and *target* networks. The online network is composed of three modules: an encoder network f(with a backbone network and a fully connected layer), a projector g and a predictor h (with the same multi-layer perceptron structure). Similarly, the target network has an encoder f' and a projector g' with different weights from the online network. As shown in Eq. 1, the weights of the target network θ' perceive an exponential moving-average [39] of the online parameters θ. We perform the moving-average after each step by target decay rate τ in Eq. 2,

$$\theta' \leftarrow \tau\theta' + (1-\tau)\theta, \tag{1}$$

$$\tau \triangleq 1 - (1-\tau_{base}) \cdot (cos(\pi s/S + 1)/2. \tag{2}$$

The exponential parameter τ_{base} is set to 0.996, s is the current training step, and S is the maximum number of training steps. In addition, a classifier head l is added after the encoder f in order to perform supervised learning. During the inference stage, only the encoder f and classifier l are applied to perform the discrimination of mask samples.

In fact, the classifier can be trained jointly with the encoder and projector networks, and achieve roughly the same results without requiring two-stage training [42]. Therefore, the CCL proposed by us is a supervised extension of the self-supervised contrastive learning. At the same time, the effective visual representation pre-learning and the downstream FAS task are completed in one stage.

C. Context Guided Dropout

In classical self-supervised contrastive learning frameworks [40]–[42], the input images x_1 and x_2 are augmented from a source image x. As a result, the similarity loss between x_1 and x_2 would decrease to a relatively low level smoothly. In contrast, our proposed CCL constructs the positive contextual pairs from separate source images, which suffer from high dissimilarity, leading to unstable convergence. Moreover, the contextual features (*e.g.*, scenes) might not always be relevant to the live/spoof cues, leading to a large similarity loss.

Inspired by the dropout operator [47], [48] to randomly discard parts of neurons during training, we propose Context Guided Dropout (CGD), which adaptively discards parts of the 'outlier' embedding features according to their similarities. For instance, given the embeddings from positive pairs, we assume that the abnormal differences between them belong to the context information. Therefore, we could automatically drop out the abnormal embeddings with huge dissimilarities after ranking their locations. For a positive n-dimensional embedding pair z_1 and p_2, we first calculate the difference vector δ_i via

$$\delta_i = |(\frac{z_1}{||z_1||_2})^2 - (\frac{p_2}{||p_2||_2})^2| \tag{3}$$

Afterward, we sort δ_i by descending sequence and record the index of the largest $p_d \cdot n$ values. Here p_d is the proportion of embedding feature channels to be discarded. And we execute this procedure in a mini-batch to determine the discarding position. Besides, the discarded embedding is scaled by a factor of $1/(1-p_d)$, which is similar to the inversed dropout method. To make CGD more adaptive, we apply a cosine decay factor to p_d during training as follows

$$p_d \leftarrow \frac{p_d}{2} \cdot (1 + cos(2\pi \frac{q_{cur}}{q})) \cdot \mathbb{1}_{q_{cur} < q/2} \tag{4}$$

where q_{cur} is the current epoch and q is the training epochs to be set. $\mathbb{1}_c \in \{0, 1\}$ is an indicator function which returns 1 if condition c is true. We also visualize the training logs in Sec. V-A. It can be seen that assembling CGD accelerates the network convergence during the early training stages, making the whole CCL training more stable.

D. Overall Loss

As CCL supervises face PAD models with live/mask binary ground truth, it is straightforward to calculate classification loss $\mathcal{L}_{cls}$ using Binary Cross Entropy (BCE) function f_{BCE}, which is described in Eq. 5. We also adopt the Eq. 6 for contrastive loss $\mathcal{L}_{con}$ calculation. To be specific, the CGD regularization f_{CGD} is firstly applied to the embedding z_1, p_2, and then Eq. 7 is applied to calculate the cosine similarity between normalized z_1 and p_2.

$$\mathcal{L}_{cls} = f_{BCE}(c_1, y_1), \tag{5}$$

$$\mathcal{L}_{con} = f_D(f_{CGD}(z_1), f_{CGD}(p_2)), \tag{6}$$

$$f_D(z_1, p_2) = (2 \cdot \mathbb{1}_{y_1 \neq y_2} - 1) \cdot < z_1, p_2 > +1. \tag{7}$$

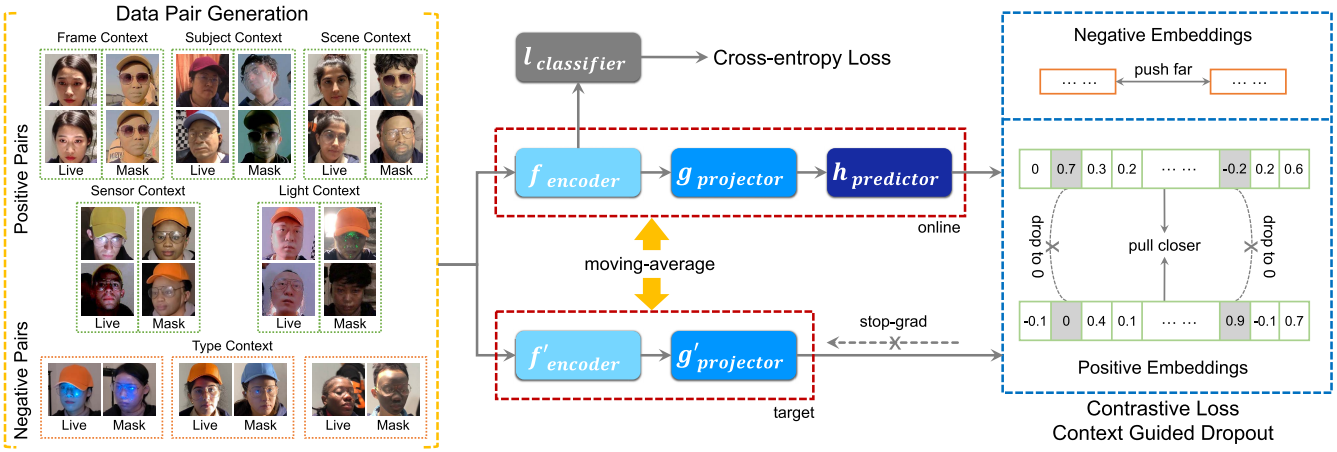

Fig. 4. The CCL framework. The left part (yellow) denotes our data pair generation manner. Each pair of images is processed by central framework twice, consisting of an online network (f, g, h), a target network without gradient backward (f', g') and a classifier header (l). The right part (blue) denotes the CGD module, the positive embeddings are pulled closer by CGD.

The overall loss $\mathcal{L}_{total}$ can be calculated by the weighted summation of $\mathcal{L}_{cls}$ and $\mathcal{L}_{con}$, *i.e.*, $\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{con} \cdot \mathcal{L}_{con}$, where λ_{con} denotes a trade-off hyper-parameter. An ablation about λ_{con} is conducted in Sec. V-A.

Algorithm 1 Training CCL

Require image set $\mathcal{X}$, label set $\mathcal{Y}$
initialize encoder f, projector g, predictor h by θ
initialize encoder f', projector g' by θ' ($\theta' = \theta$)
while not end of training
  sample image pairs $(x_1, x_2) \subseteq \mathcal{X}^2$, corresponding y_1, y_2
  compute online branch and classification result on x_1 by
    $z_1 = h \circ g \circ f(x_1)$, $c_1 = l \circ f(x_1)$
  compute target branch on x_2 by
    $p_2 = g' \circ f'(x_2)$
  do CGD procedure on z_1, p_2 by
    $z_1 \leftarrow f_{CGD}(z_1)$, $p_2 \leftarrow f_{CGD}(p_2)$
  do stop gradient on p_2
  compute $\mathcal{L}_{cls}$, $\mathcal{L}_{con}$, $\mathcal{L}_{total}$ by Eq. 5, 6, 7
  compute gradient $\Delta\theta$ and update θ by
    $\Delta\theta = backward(\mathcal{L}_{total})$
    $\theta \leftarrow \theta - learning\_rate \cdot \Delta\theta$
  update θ' by Eq. 1, update p_d by Eq. 4
until converged

notation: $g \circ f(x)$ represents the composite function of g and f

V. Experiments

In this section, we conduct a series of experiments on the HiFiMask and other widely used face PAD datasets.

Datasets & Protocols: Four datasets are used in our experiments: WMCA [17], CASIA-SURF 3DMask (briefly named 3DMask) [22], HKBU-MARsV2 (briefly named MARsV2) [26], and the proposed HiFiMask. We perform Intra Testing on HiFiMask and WMCA datasets with the 'seen' and 'unseen' protocols, and study Cross Testing performance on 3DMask and MARsV2 datasets when training on HiFiMask.

Performance Metrics: In HiFiMask and WMCA datasets, Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and ACER [49] are used for performance evaluation. The ACER on the testing set is determined by the Equal Error Rate (EER) and BPCER = 1% thresholds on development sets for HiFiMask and WMCA, respectively. In the Cross Testing experiments on 3DMask and MARsV2 datasets, Half Total Error Rate (HTER) [50] and Area Under Curve (AUC) are adopted as evaluation metrics.

Architecture: We use a universal network ResNet50 [24], Aux.(Depth) [4], CDCN [51] and ViT [52] with varying dimension of the last layer as backbones, and report their results as baselines. The last full connected(FC) layer in encoder f and f' is set from 2048 dimensions to 128 dimensions for ResNet50 backbone. As in SimCLR [40], a projector g and g' is introduced behind encoder. The projector consists of a hidden FC with 512 dimension output followed by batch normalization [53] and ReLU layers. The last layer in projector is FC only with output dimension 128. After the projector, predictor h has the same architecture as projector.

Optimization: Unless specified, we adopt SGD with weight decay 0.0001 and momentum 0.9 for the model training. The total batch size is 256 on eight 2080Ti GPUs. The learning rate starts with 0.01, and decays by $\gamma = 0.2$ once the number of epoch reaches one of the milestones. Models are trained for 30 epochs with milestones in *15, 21, 26.* For ViT backbone, we used setting from paper, adopting AdamW as optimization and learning rate of 0.0001.

A. Ablation Study

Here we conduct ablation experiments to verify the contributions of each module of the proposed CCL on Protocol 1 of the HiFiMask dataset.

1) Effect of Architectures: As shown in Tab. III, we select ResNet50 [24] as baseline and equip it with five different contrastive-based learning strategies for comparison, such as Contrastive Loss [36], Triplet Contrastive Loss [37], SimSiam [42], BYOL [41] and supervised contrastive learning method SupCon [43] (SC for abbreviation). The results show that contrastive-based learning is more suitable for mining the discrepancy between live face and mask material than

TABLE III

THE ABLATION RESULTS ON THE PROTOCOL 1 OF HIFIMASK

Method	APCER(%)	BPCER(%)	ACER(%)
ResNet50 [24]	3.7	5.7	4.7
ResNet50 w/ Contrastive Loss [36]	3.2	4.2	3.7
ResNet50 w/ Triplet Loss [37]	3.5	4.9	4.2
ResNet50 w/ SimSiam [42]	2.4	5.7	4.0
ResNet50 w/ BYOL [41]	**1.7**	4.0	2.9
ResNet50 w/ SC [43]	3.2	3.2	3.2
CCL w/o CGD	1.9	3.3	2.6
CCL w/ dropout	2.5	3.5	3.0
CCL w/ reverse CGD	2.2	3.5	2.8
CCL w/ BOBE CGD	2.0	3.4	2.7
CCL	1.8	**3.0**	**2.4**

the vanilla ResNet50, with the ACER improvement from 4.7% to 2.9%.

In terms of self-supervised contrastive learning approaches, SimSiam and BYOL achieve 4.0% and 2.9% ACER on Protocol 1 of HiFiMask, respectively. It is clear that BYOL outperforms SimSiam by a large margin, indicating the importance of moving-average when updating network parameters. Based on the moving-average mechanism, CCL further decreases the ACER by 0.5% compared with BYOL. This is because CCL is able to efficiently exploit the elaborated contextual pairs for fine-grained feature representation while BYOL only considers the simple augmented views. In Tab. III, we also compare CCL with the recent supervised contrastive learning method SC. The CCL achieves better performance (reducing 0.8% ACER) than SC, which indicates the advances of our context-aware pair generation and CGD strategies for 3D mask PAD task.

In summary, the proposed CCL extends the self-supervised contrastive approach to the supervised setting and pulls together the clusters of points that belong to the same class while simultaneously pushes apart clusters of samples from different classes in the embedding space. It allows us to effectively leverage label information to mine the differences between face skin and different kinds of mask materials.

2) Effect of Data Pair Generation: An effective Data Pair Generation can accelerate the model convergence and guide the network to mine liveness-related features. As shown in Tab. IV, we performed such ablation experiments by combining different kinds of image pairs. After adding Sensor, Light, Scene, Type and Subject image pairs sequentially, our training set becomes larger with an increasing numbers of contrastive categories. As a result, the ACER is decreased from 3.9% to 2.4%, which shows a significant effect of the proposed contrastive patterns in our experiment. One can observe that the data pair generated under different patterns has varied performances.

3) Effect of Data Organization: As illustrated in Tab. V, we take context such as identity, mask, lighting, imaging sensor, and frame information into consideration to generate data pairs with 6 different patterns. We practice these six patterns to perform experiments on protocol 1 of HiFiMask, respectively. When only negative pairs are used, it is obtained an ACER of 43.7%. The main reason is that no positive samples are available in the training set makes the model convergence difficult. The mixed-use of multiple positive and negative pairs can optimize the performance of the model.

4) Effect of CGD: The proposed CGD module can accelerate the model convergence in the early stage of training and alleviate the interference of useless information to the model. As shown in Tab. III, if the CGD is removed, the performance of three indicators decreases, with an APCER, BPCER and ACER increasing from 1.8%, 3.0%, and 2.4% to 1.9%, 3.3%, and 2.6%, respectively.

TABLE IV

ABLATION STUDY OF COMBINING DIFFERENT IMAGE PAIRS

Pat.	Subject	Type	Scene	Light	Sensor	Frame	Pair	ACER(%)
1					✓	✓	pos	3.9
2				✓	✓	✓	pos	3.2
3			✓	✓	✓	✓	pos	3.0
4		✓	✓	✓	✓	✓	pos&neg	2.9
5	✓	✓	✓	✓	✓	✓	pos&neg	2.4

TABLE V

PATTERNS TO ORGANIZE IMAGE PAIRS BY ADJUSTING ATTRIBUTES

Pat.	Subject	Type	Scene	Light	Sensor	Frame	Pair	ACER(%)
1						✓	pos	3.6
2					✓	✓	pos	3.9
3				✓		✓	pos	4.0
4			✓			✓	pos	3.2
5		✓				✓	neg	43.7
6	✓					✓	pos	3.9
1-6	✓	✓	✓	✓	✓	✓	pos&neg	2.4

In the proposed CGD, we aim to manually discard the most dissimilar neurons in the embedding by finding their location for a positive embedding pair. In order to verify the effectiveness of removing embedding points, we introduce two variants for comparison: 'reverse' CGD and 'BOBE'(Break Out Both Ends) CGD. The first one removes the embedding points in the opposite way to the CGD. The second one removes the most similar and the least similar embedding points at the same time. Their performance is decreased, as expected, to 2.8% and 2.7% for ACER. Compared with random dropout, our CGD discards the abnormal neurons in the embeddings more pertinently, such as the huge dissimilarities caused by different source images in the positive sample pairs. The experimental results verify the correctness of our CGD design concept. For example, by comparing 'CCL w/ dropout' and 'CCL' in Tab. III, our CGD has a performance gain of 0.6% on ACER. To study the impact of the Context Guided Dropout (CGD) during the training stage, we also draw the loss curve shown in Fig. 5. Intuitively, It can be seen that the red line (with CGD) drops faster than the blue line (without CGD) in the early training stage, which proves that CGD accelerates the network convergence during the early training stages making the whole CCL training more stable.

5) Effect of Parameter λ_{con} and p_d: Derived from the popular dropout operator, we can foresee that the effect of CGD would be affected by the probability of randomly abandoning neurons during training time. Also, factor λ_{con} in Sec. 4.4 controls the relative importance of the CCL loss $\mathcal{L}_{con}$ and BCE loss $\mathcal{L}_{cls}$ in overall loss $\mathcal{L}_{tolal}$. Therefore, different λ_{con} will affect the performance gains. In order to eliminate the stochastic behavior of the algorithms and randomness of the experiments, obtain the best values of p_d and λ_{con}, we repeated the experiment three times for each systematic change, taking the mean value of the corresponding ACER as the result. As shown in Fig. 6 (b), when the probability p_d increases from 0 to 25%, the performance increases from 2.78 to 2.66 for ACER. However, as the p_d continues to increase, some functional neurons are lost, leading to performance degradation. Similarly, see from the Fig. 6 (a),

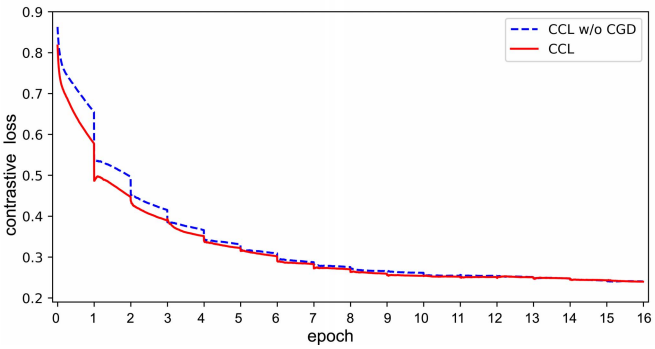

Fig. 5. The contrastive loss during training stage. (Red line: CCL with CGD; Blue line: CCL without CGD.) **Best viewed in color.**

when the proportion of $\mathcal{L}_{con}$ is 0.7, the model reaches the optimal performance in Protocol 1 of HiFiMask, that is, the ACER reaches the minimum value of 2.50%. Based on the above experiments, we fixed the probability p_d at 15% and λ_{con} at 0.7 in subsequent experiments.

6) Effect of rPPG Signal Recovery With Periodic Lights: Here we show a simple example to illustrate the challenges on the proposed HiFiMask about the rPPG recovery. We follow the approach GrPPG [25] to firstly track the facial region of interests (ROI), and then the intensity values within the ROI from each color channel are averaged to form the rPPG signals. As illustrated in the top right sub-figure of Fig. 7(a), the rPPG signals extracted from mask attack without periodic light are quite noisy, indicating the weak heartbeat-derived liveness clues. In contrast, it can be seen from Fig. 7(b) that, under the scenario with periodic light within [0.7,4]Hz, the recovered rPPG signals are with rich periodicity. We can also see from the bottom right sub-figures of Fig. 7(a)(b) that the power spectral density (PSD) distributions of the extracted rPPG signals are clear to show periodicity/liveness evidences. Thus, it would mislead the rPPG-based attack detector for incorrect decision, i.e., treating the mask face with periodic light as a live face. Besides qualitative analysis, we also conduct two experiments on the Protocol 1 using the subsets w/o and w/ period lights, respectively. As for the former case, GrPPG [25] can achieve 19.6% EER. However, when extracting rPPG features under the challenging illumination variation, the performance (49.1% EER) drops sharply.

In the HiFiMask dataset, as the temporal light conditions are diverse for both the bonafides and 3D mask attacks, it is not easily to use current rPPG approaches for robust PAD. Moreover, the dynamic background with pedestrian movement makes it more challenging for global noises compensation used in recent rPPG methods [21], [34]. As there are finger-contacted BVP signals as well as instantaneous heart rate values as groundtruth for the bonafides, in the future one possible direction is to design a robust rPPG extractor and liveness detector even under complex temporal light interference.

B. Intra Testing

1) Intra Testing on HiFiMask: As shown in Tab. VII, without using the proposed CCL, the ViT with pre-trained achieves the lowest ACER (*i.e.*, Protocol 1: 3.4%, Protocol 2: 11.4%) when compared with ResNet50 and CDCN. We integrated

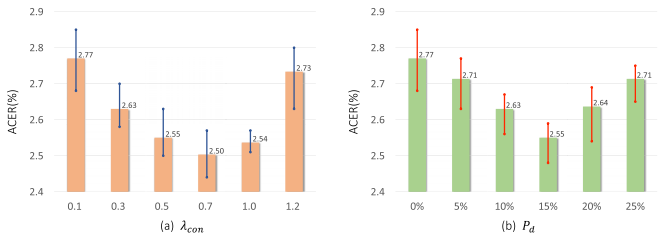

Fig. 6. Ablation studies on (a) CCL and(b) CGD on the Protocol 1 of HiFiMask.

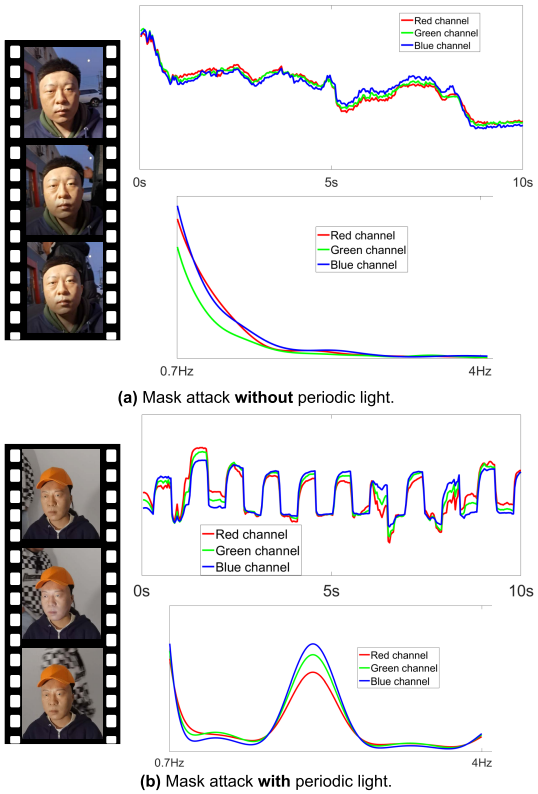

Fig. 7. Visualization of video samples and their extracted rPPG signal and power spectral density (PSD). In both cases, i.e., without (a) and with (b) external periodic light, the top right sub-figures illustrate the rPPG signal in time domain while the bottom right ones show their PSD in frequency domain. It can be seen from (b) that with period light, mask attacks could also contain pseudo 'live' pulse clues. **Best viewed in color**.

the three backbones into CCL framework, and the performances are improved consistently. Specifically, the ACERs of ResNet50, Aux.(Depth), CDCN are decreased from 4.7%, 3.4%, 3.6% to 2.4%, 2.6%, 3.1% on Protocol 1, and from 18.5%, 11.4%, 14.7% to 16.7%, 10.7%, 12.7% on Protocol 2, respectively.

In protocol 2, ACER of ResNet50, Aux.(Depth) and CDCN on Protocol 2_1 are 34.0%, 24.1% and 30.9%, respectively. The reason lies in the appearance of transparent mask varies from the plaster and resin masks (see Fig. 2). Notably, our method still steadily reduce the ACER of the three benchmarks (32.5%, 22.6% and 26.6%, respectively).

The ACER of ResNet50 on Protocol 3 is 20.9%, higher than 18.5% on Protocol 2, meaning Protocol 3 is more challenging to face PAD. The reason may be that the larger 'openset' between training and testing sets demands for a higher representation and generalization capabilities algorithm. In this case, the proposed CCL still obtains better performance,

TABLE VI

COMPARISON OF THE RESULTS OF PROTOCOLS 'SEEN' AND 'UNSEEN' ON WMCA. THE VALUES ACER(%) REPORTED ON TESTING SETS ARE OBTAINED WITH THRESHOLDS COMPUTED FOR BPCER=1% ON DEVELOPMENT SETS. 'RGB-D' DENOTES USING BOTH RGB AND DEPTH INPUTS. METHODS WITH '*' DENOTE USING PRE-TRAINED MODEL

Modality	Method	Seen	Unseen							
			Flexiblemask	Replay	Fakehead	Prints	Glasses	Papermask	Rigidmask	Mean±Std
RGB-D	MC-PixBiS	1.8	49.7	3.7	0.7	0.1	16.0	0.2	3.4	10.5±16.7
	MCCNN-OCCL-GMM	3.3	22.8	31.4	1.9	30.0	50.0	4.8	18.3	22.74±15.3
	MC-ResNetDLAS	4.2	33.3	38.5	49.6	3.80	41.0	47.0	20.6	33.4±14.9
RGB	ResNet50	40.93	14.48	15.69	38.00	32.71	27.33	20.14	30.22	25.51±8.95
	ResNet50 w/ CCL	30.69	**4.76**	15.37	24.67	19.03	16.80	9.51	17.62	**15.39±6.51**
	Aux.(Depth)	42.67	13.23	12.52	47.33	32.18	23.69	13.92	40.43	26.19±14.13
	Aux.(Depth) w/ CCL	30.62	7.41	12.76	40.00	16.11	**10.17**	11.67	27.32	17.92±11.66
	CDCN	38.41	12.10	**8.69**	42.67	30.07	11.67	11.87	30.38	21.06±13.17
	CDCN w/ CCL	27.14	7.18	11.79	21.82	20.53	35.13	18.91	15.10	18.64±8.91
	ViT*	26.62	24.61	24.43	6.44	10.74	50.00	**1.16**	8.15	17.93±16.71
	ViT* w/ CCL	**24.33**	19.09	30.70	**6.00**	**9.96**	45.50	**1.16**	**4.07**	16.64±16.29

TABLE VII

THE RESULTS OF INTRA TESTING ON THE HIFIMASK. METHODS WITH '*' DENOTE USING PRE-TRAINED MODEL

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	ResNet50	3.7	5.7	4.7
	Aux.(Depth)	4.9	1.8	3.4
	CDCN	3.3	3.9	3.6
	ViT*	2.4	**1.5**	1.9
	ResNet50 w/ CCL	1.8	3.0	2.4
	Aux.(Depth) w/ CCL	2.1	3.1	2.6
	CDCN w/ CCL	3.0	3.3	3.1
	ViT* w/ CCL	**1.0**	1.9	**1.4**
2	ResNet50	22.4±15.3	14.6±6.7	18.5±11.0
	Aux.(Depth)	11.1±9.4	11.2±9.8	11.2±9.0
	CDCN	12.6±7.3	16.8±15.6	14.7±11.4
	ViT*	12.2±10.3	12.9±11.2	12.5±10.7
	ResNet50 w/ CCL	16.7±11.2	16.7±12.4	16.7±11.2
	Aux.(Depth) w/ CCL	**10.7±7.5**	**10.7±9.4**	**10.7±8.4**
	CDCN w/ CCL	13.3±10.6	12.2±9.2	12.7±9.8
	ViT w/ CCL	14.4±13.6	10.3±7.8	12.4±10.7
3	ResNet50	13.5	28.3	20.9
	Aux.(Depth)	9.6	16.2	12.9
	CDCN	20.8	12.5	16.7
	ViT*	6.4	2.5	4.5
	ResNet50 w/ CCL	13.8	23.4	18.6
	Aux.(Depth) w/ CCL	8.2	12.7	10.5
	CDCN w/ CCL	15.4	13.2	14.3
	ViT* w/ CCL	**4.6**	**2.5**	**3.5**

TABLE VIII

CROSS-TESTING RESULTS ON DIFFERENT DATASETS. METHODS WITH '*' DENOTE USING PRE-TRAINED MODEL

Method	3DMask to MARsV2		MARsV2 to 3DMask	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
ResNet50	37.96	67.05	43.69	59.03
CDCN	45.20	56.13	32.56	73.60
Aux.(Depth)	44.24	57.05	43.19	60.08
ViT*	34.83	59.86	50.00	44.39
ResNet50 w/ CCL	41.49	61.22	40.32	64.52
CDCN w/ CCL	42.45	56.26	38.20	66.84
Aux.(Depth) w/ CCL	43.01	57.66	36.27	67.32
ViT* w/ CCL	27.12	74.03	39.42	61.75
Method	HiFiMask to MARsV2		HiFiMask to 3DMask	
	HTER(%)↓	AUC(%)↑	HTER(%)↓	AUC(%)↑
ResNet50	20.61	86.87	30.35	76.36
CDCN	16.56	90.81	17.28	89.94
Aux.(Depth)	9.31	96.31	16.11	91.32
ViT*	9.82	96.72	17.14	90.74
ResNet50 w/ CCL	14.04	90.66	25.43	81.99
CDCN w/ CCL	15.66	92.25	**13.97**	**93.26**
Aux.(Depth) w/ CCL	6.98	98.42	16.33	90.67
ViT* w/ CCL	**5.72**	**98.69**	16.12	92.03

with a decrease of 2.3%, 2.4% and 2.4%, respectively. When training on ViT with/without CCL, we used pre-trained model from ImageNet.

The results from Tab. VII show that the proposed CCL provides good generalization with different backbones. One could thus expect that considering newly emergent networks within the CCL framework could potentially improve current performance.

2) Intra Testing on WMCA: The experimental results of protocols 'seen' and 'unseen' on WMCA [17] are shown in Tab. VI. The results in the third column show that all four widely used face PAD backbones (*i.e.*, ResNet50, Aux.(Depth), CDCN, and ViT) assembled with CCL can achieve significantly lower ACER values (with a decrease of 10.24%, 12.05%, 11.27%, and 2.29% respectively). This indicates that the proposed CCL effectively leverages the context cues (e.g., rich attack types) to learn more discriminative features. In terms of the 'unseen' protocol, we follow the same setting as [17]. We can draw similar conclusions that the proposed CCL benefits the unseen attacks for all four backbones. To be specific, compared with the vanilla ResNet50, Aux.(Depth), CDCN, and ViT, the counterparts with CCL can reduce the values of ACER to 15.39%, 17.92%, 18.64%, and 16.64%, respectively. It is worth noting that the performance of 'ViT' and 'ViT w/ CCL' against 'Replay' and 'Glasses' are poor. We analyze that Transformer identifies sample from the dependency between face tokens. These two kinds of spoofing clues only appear in some patch tokens, and the relationship with other tokens is difficult to model.

C. Cross Testing

1) Cross Testing of Datasets: As shown in Tab. VIII, we first train a model on 3DMask dataset and directly test on MARsV2 dataset, and repeat the procedure by exchanging the two datasets. Then, we train a model on proposed HiFiMask dataset and test on MARsV2 and 3DMask respectively.

Comparing the results in Tab. VIII, while the training set is our proposed HiFiMask, the HTER values are relatively lower and the AUC values higher. For example, HTER values are 20.61%, 16.56%, 9.31% and 9.82% when training with HiFiMask and testing with MARsV2 on four architectures above. These values increase to 37.96%, 45.20%, 44.24% and 34.83% if the training set is set to 3DMask. What's more, there is a significant drop of AUC values from 86.87%, 90.81%, 96.31% and 96.72% to 67.05%, 56.13%, 57.05% and 59.86%. This phenomenon can be observed as well if testing set is

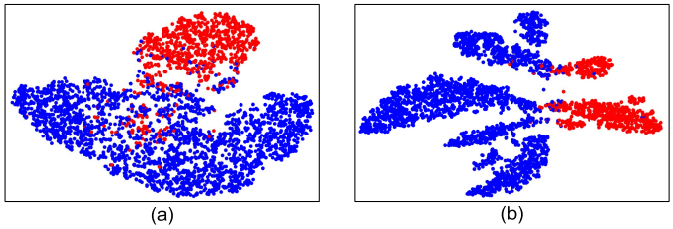

Fig. 8. Feature distribution comparison on Protocol 1 of HiFiMask using t-SNE [54]: (a) ResNet50, (b) the proposed CCL. The points with different colors denote features from different classes (red: real faces; blue: mask samples). **Best viewed in color.**

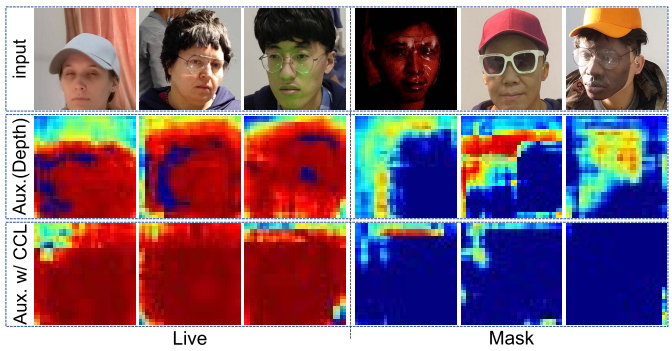

Fig. 9. Visualization of several samples in HiFiMask. The maps are performed from Aux.(Depth) in the second row and Aux. w/ CCL in the third row (Red color: the higher the better for live samples; Blue color: the higher the better for mask samples). **Best viewed in color.**

3DMask dataset. From this, we can conclude that our proposed HiFiMask is a well-distributed mask dataset, which has a better generalization than MARsV2 and 3DMask.

To further evaluate the generalization of CCL, we report the results of each backbone equipped with our CCL framework. As shown in Tab. VIII, compared with bare backbones, the proposed CCL obtains better performance in most configurations. Such as ViT with CCL achieves the best performance under "HiFiMask to MARsV2" and CDCN with CCL achieves the best performance under "HiFiMask to 3DMask".

In Tab. VIII, one can observe that CCL failed to further improve the performance of ResNet50 in "3DMask to MARsV2", CDCN in "MARsV2 to 3DMask" and Aux.(Depth) in "HiFiMask to 3DMask". The main reason might be that MARsV2 and 3DMask has only a limited attack types, which inhibits the advantage of our Data Pair Generation strategy to generate abundant context information.

D. Analysis and Visualization

In this section, we further visually analyze the CCL's ability to distinguish different material features. As shown in Fig. 8, we compare the features learned by ResNet50 and CCL on HiFiMask (Protocol 1). Compared with ResNet50, the proposed CCL can well distinguish real faces from mask samples. In addition, we visualize the regression results (heatmaps) by Aux.(Depth) and Aux.(Depth) with CCL in Fig. 9. We can see that Aux.(Depth) fails to distinguish the complex mask samples in the last two columns while the proposed Aux.(Depth) with CCL provides more correct predictions. It demonstrates the better discriminative representation capacity of the proposed CCL.

VI. Conclusion

In this paper, we released a large-scale HiFiMask dataset with three challenging protocols. We hope it will push cutting-edge research in 3D Mask face PAD. Besides, we proposed a novel CCL framework to learn discriminability by leveraging rich contexts among pairs of live faces and mask attacks. Finally we conducted a comprehensive set of experiments on both HiFiMask and other three 3D mask datasets, verifying the significance of the proposed dataset and method.

VII. Acknowledgment

Ajian Liu is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China (e-mail: ajianliu92@gmail.com).

Chenxu Zhao, Anyang Su, and Xing Liu are with the Mininglamp Academy of Sciences, Mininglamp Technology, Beijing 10010, China (e-mail: zhaochenxu@mininglamp.com).

Zitong Yu is with the Faculty of Information Technology and Electrical Engineering, University of Oulu, 90570 Oulu, Finland (e-mail: zitong.yu@oulu.fi).

Jun Wan is with the Faculty of Innovation Engineering, Macau University of Science and Technology (MUST), Macau, also with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100190, China (e-mail: jun.wan@ia.ac.cn).

Zichang Tan and Guodong Guo are with the Baidu Research and National Engineering Laboratory for Deep Learning Technology and Application, Institute of Deep Learning, Beijing 100081, China (e-mail: tanzichang@baidu.com; guoguodong01@baidu.com).

Sergio Escalera is with the Computer Vision Centre (CVC), Universitat de Barcelona (UB), 08007 Barcelona, Spain (e-mail: sergio@maia.ub.es).

Junliang Xing is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100190, China (e-mail: jlxing@tsinghua.edu.cn).

Yanyan Liang and Du Zhang are with the Faculty of Innovation Engineering, Macau University of Science and Technology (M.U.S.T), Macau, China (e-mail: yyliang@must.edu.mo; duzhang@must.edu.mo).

Zhen Lei is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100190, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China (e-mail: zlei@nlpr.ia.ac.cn).

Stan Z. Li is with the AI Lab, School of Engineer, Westlake University, Hangzhou 310024, China (e-mail: stan.zq.li@westlake.edu.cn).

REFERENCES

[1] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. ICB*, Mar. 2012, pp. 26–31.

[2] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. BIOSIG*, 2012, pp. 1-7.

[3] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. BTAS*, Sep. 2013, pp. 1–8.

[4] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. CVPR*, Jun. 2018, pp. 389–398.

[5] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Proc. ICB*, 2019, pp. 1–8.

[6] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, "Face anti-spoofing with human material perception," in *Proc. ECCV*, 2020, pp. 557–575.

[7] K.-Y. Zhang *et al.*, "Face anti-spoofing via disentangled representation learning," in *Proc. ECCV*, 2020, pp. 641–657.

[8] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 406–422.

[9] B. Yang, J. Zhang, Z. Yin, and J. Shao, "Few-shot domain expansion for face anti-spoofing," 2021, *arXiv:2106.14162*.

[10] Z. Chen *et al.*, "Generalizable representation learning for mixture domain face anti-spoofing," 2021, *arXiv:2105.02453*.

[11] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. FGR*, May 2017, pp. 612–618.

[12] S. Zhang *et al.*, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proc. CVPR*, Jun. 2019, pp. 919–928.

[13] Y. Zhang *et al.*, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Proc. ECCV*, 2020, pp. 70–85.

[14] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proc. BTAS*, Sep. 2013, pp. 1–8.

[15] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 7, pp. 1794–1809, Jul. 2018.

[16] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proc. CVPR*, 2019, pp. 4680–4689.

[17] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 42–55, 2020.

[18] G. Heusch, A. George, D. Geissbuhler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 4, pp. 399–409, Oct. 2020.

[19] N. Kose and J.-L. Dugelay, "Mask spoofing in face recognition and countermeasures," *Image Vis. Comput.*, vol. 32, no. 10, pp. 779–789, Oct. 2014.

[20] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral SWIR imaging," in *Proc. ICB*, Jun. 2016, pp. 1–8.

[21] S.-Q. Liu, X. Lan, and P. C. Yuen, "Remote photoplethysmography correspondence feature for 3D mask face presentation attack detection," in *Proc. ECCV*, 2018, pp. 558–573.

[22] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "NAS-FAS: Static-dynamic central difference network search for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3005–3023, Sep. 2021.

[23] A. George and S. Marcel, "Cross modal focal loss for RGBD face anti-spoofing," in *Proc. CVPR*, Jun. 2021, pp. 7882–7891.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[25] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *Proc. ICPR*, Dec. 2016, pp. 4244–4249.

[26] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3D mask face anti-spoofing database with real world variations," in *Proc. CVPRW*, Jun. 2016, pp. 100–106.

[27] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, vol. 5, no. 2, pp. 83–91, Jun. 2016.

[28] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1713–1723, Jul. 2017.

[29] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proc. CVPRW*, Jul. 2017, pp. 81–89.

[30] S. Bhattacharjee and S. Marcel, "What you can't see can help you—Extended-range imaging for 3D-mask presentation attack detection," in *Proc. BIOSIG*, Sep. 2017, pp. 1–7.

[31] S. Jia, G. Guo, and Z. Xu, "A survey on 3D mask presentation attack detection and countermeasures," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107032.

[32] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3D face anti-spoofing with factorized bilinear coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4031–4045, Oct. 2021.

[33] B. Lin, X. Li, Z. Yu, and G. Zhao, "Face liveness detection by rPPG features and contextual patch-based CNN," in *Proc. ICBEA*, 2019, pp. 61–68.

[34] S.-Q. Liu, X. Lan, and P. C. Yuen, "Multi-channel remote photoplethysmography correspondence feature for 3D mask face presentation attack detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2683–2696, 2021.

[35] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, "3D mask face anti-spoofing with remote photoplethysmography," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 85–100.

[36] H. Hao, M. Pei, and M. Zhao, "Face liveness detection based on client identity using Siamese network," in *Proc. PRCV*. Springer, 2019, pp. 172–180.

[37] L. Li, Z. Xia, X. Jiang, F. Roli, and X. Feng, "CompactNet: Learning a compact space for face presentation attack detection," *Neurocomputing*, vol. 409, pp. 191–207, Oct. 2020.

[38] G. Wang, H. Han, S. Shan, and X. Chen, "Improving cross-database face presentation attack detection via adversarial domain adaptation," in *Proc. ICB*, Jun. 2019, pp. 1–8.

[39] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, Jun. 2020, pp. 9729–9738.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[41] J.-B. Grill *et al.*, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. NeurIPS*, 2020, pp. 21271–21284.

[42] X. Chen and K. He, "Exploring simple Siamese representation learning," 2020, *arXiv:2011.10566*.

[43] P. Khosla *et al.*, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.

[44] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[45] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "FaceX-zoo: A PyTorch toolbox for face recognition," 2021, *arXiv:2101.04407*.

[46] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, Jun. 2019, pp. 4690–4699.

[47] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. CVPR*, Jun. 2019, pp. 2219–2228.

[48] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1249–1258.

[49] International Organization for Standardization. (2016). *ISO/IEC JTC 1/SC 37 Biometrics: Information Technology Biometric Presentation Attack Detection Part 1: Framework*. [Online]. Available: https://www.iso.org/obp/ui/iso

[50] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Odyssey, Speaker Lang. Recognit. Workshop*, 2004, pp. 1–8.

[51] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. CVPR*, Jun. 2020, pp. 5295–5305.

[52] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[54] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, Nov. 2008.